Hwk #1: Regression and Classification
Biostat 216: Machine Learning in R for the Biomedical Sciences

For this homework we will use NHANES data that exists in a package for R.

NHANES consists of survey data collected by the US National Center for Health Statistics (NCHS) which has conducted a series of health and nutrition surveys since the early 1960's. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination center (MEC).

Note that there is a an NHANES warning that comes from their website.
"For NHANES datasets, the use of sampling weights and sample design variables is recommended for all analyses because the sample design is a clustered design and incorporates differential probabilities of selection. If you fail to account for the sampling parameters, you may obtain biased estimates and overstate significance levels."

We are going to ignore this warning and just apply our analyses to the data as if they were randomly sampled! There is a resampled dataset in the package that is re-sampled to represent a more random sample, but we will ignore that here. We will be using the data called `NHANESraw`.

For homework questions below that correspond to tasks in R please provide the code that you used along with relevant output. You can submit your homework based `RMarkdown` if you are comfortable using that, or you can simply cut and paste into this word document.

Note that when you are asked to comment in questions below just one or two sentences should suffice in each case.

Data Preparation

1. Install the package `NHANES` into `R`, load the NHANES package, and then run the command `data(NHANES)` which will load the NHANES data. Type `?NHANES` and read about the dataset.
2. Make an object `nhanes` that is a subset version `NHANESraw` that does not include any missing data for `Diabetes`, `BPSysAve`, `BPDiaAve`, or `Age` using the following (or your own alterative approach):

```
nhanes <- subset(NHANESraw,subset=is.na(Diabetes)==FALSE &
is.na(BPSysAve)==FALSE & is.na(BPDiaAve)==FALSE & is.na(Age)==FALSE)
```

3. (1 point) Plot `BPSysAve` against `BPDiaAve`. Comment on what is going on.

4. Further subset the data such the observations with `BPDiaAve` equal to zero are removed.
5. Make an object `nhanes09` that is a subset of `nhanes` to only the 2009_10 data. This will be your training dataset. Also make an object `nhanes11` that is a subset of `nhanes` to only the 2011_12 data. This will be your test dataset.

Linear regression

6. (1 point) Plot `BPSysAve` against `BPDiaAve` for the training data.
7. (2 points) Fit a linear model in the training data with `BPSysAve` as outcome and `BPDiaAve` as the single predictor.
8. (1 point) Use the `summary` command to examine the resulting fitted model.
9. (1 point) Generate 95% confidence intervals for the parameters of the fitted model.
10. (1 point) Also, generate 99% confidence intervals for the parameters of the fitted model.
11. (2 points) Comment on the difference between the 95% and 99% confidence intervals and whether or not the difference is what you would expect.
12. (3 points) Now fit models with quadratic and cubic terms in the predictor `BPDiaAve` in addition to the linear term. Look at the output of each model with summary and generate 95% confidence intervals for each.
13. (2 points) Add the linear, quadratic and cubic fit lines on to the plot in different colors.
14. (1 point) Which would be your preferred model based on the visual fits?
15. (3 points) Perform an `anova` test comparing the 3 models. Does the result seem in line with what you were expecting from the visual fits? Why/why not?
16. (1 point) Now plot `BPSysAve` against `BPDiaAve` for the test data.
17. (1 point) Add the training data model fit lines to the test data plot.
18. (3 points) Does this change your opinion at all about which is the best fit? Why/why not?

Smoothing kernels:

19. (3 points) Fit a nearest neighbors curve with `ksmooth` using the `"normal"` `kernel` to the `nhanes09` data with bandwidths of 3, 5 and 10.
20. (2 points) Add these curves to your preferred visually fitted curve from linear regression, again using different color lines.
21. (2 points) Which is your preferred curve to represent the data and why?